

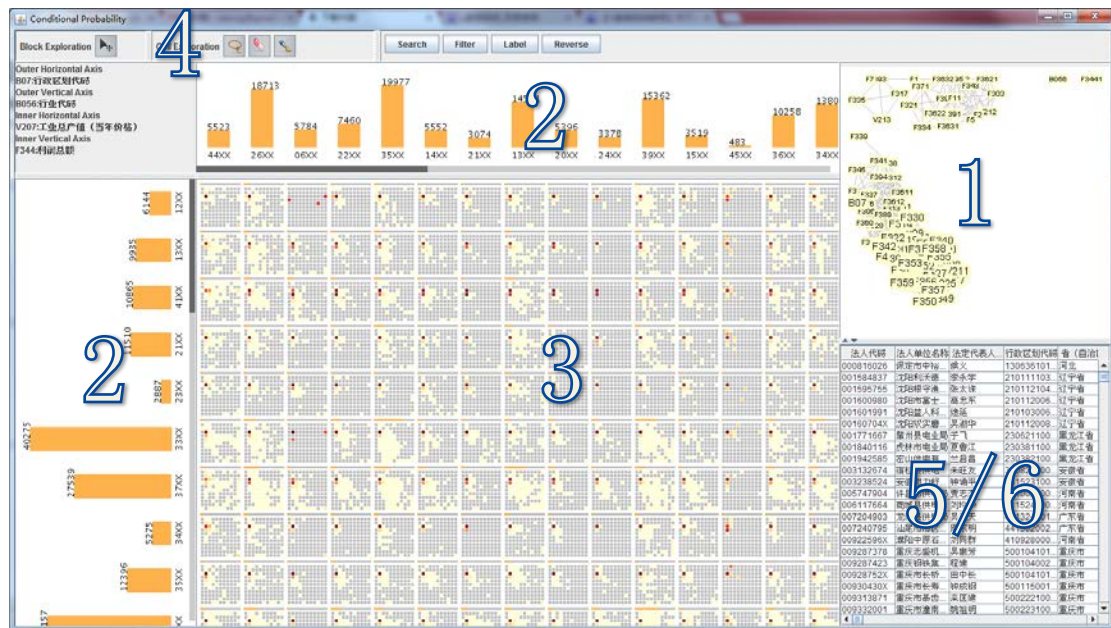
本周周报

解聪

2013.12.9-2013.12.15

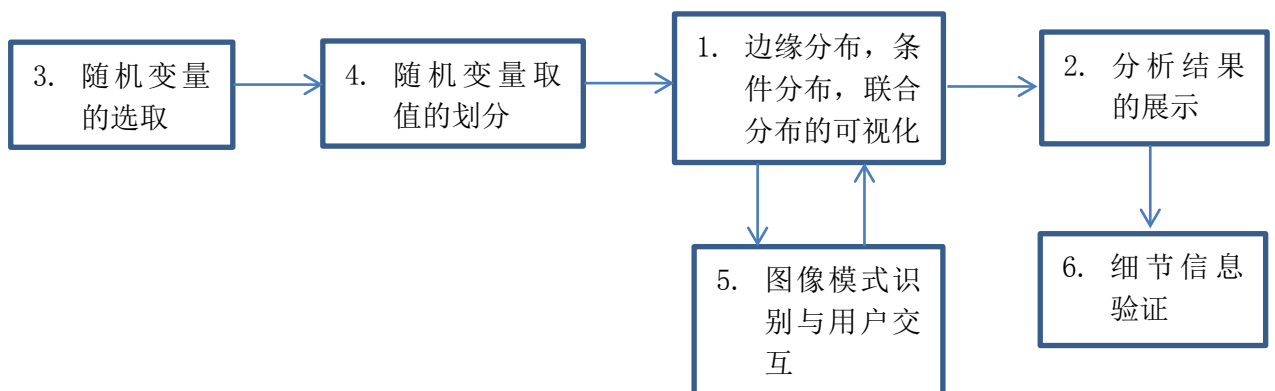
本周工作

1. 毕业论文与专利的书写。
 2. 继续完善离散随机变量的探索系统。
- 本周基本上完成了离散变量分布的系统框架。



在企业数据的探索过程中，很多地方需要手动调整一下，比如维度的划分，颜色的映射等。虽然这些都是一些琐碎的工作，但是对于可视化效果影响很大。这也是接下去系统要加入的功能。

3. 离散变量分布探索的流程。



1. 随机变量的选取:

首先由计算各变量的的信息与个变量之间的关联（计算信息熵与互信息）

变量关系的可视化，用户进一步选取感兴趣的变量（变量关系图的展示与交互）

2. 连续随机变量的数值划分：

划分准则是使得划分后的条件概率熵尽量小，转化为求解一个非线性多变量函数的最小值。可以使用 Nelder-Mead Algorithm 来做。使用 Matlab 中 Optimization Toolbox 里的函数比较方便。

3. 离散随机变量分布的可视化：

使用类似于 mosaic 图划分的矩阵展示：边缘分布（柱状图），联合分布（条形图）以及条件分布（矩阵图）。

4. 图像模式的识别与可视化交互

分析师关心的任务可能包括：

1. 随机变量取值之间的相关性（矩阵行列的重排列，筛选，聚类）。
2. 条件概率密度函数的局部相似性的查找（距离与相关系数的计算，矩阵中每一块的点击拖动划选，框选，橡皮擦等各种交互）。
3. 条件概率密度函数相似性的修正（距离尺度学习）。
4. 模糊理论与贝叶斯概率的相关知识。（有待探索）
5. 数据的标记与规则的导出。

存在的图像模式包括：

对于一个条件分布：全局分布的特征与局部分布的特征。

对于两个或多个条件分部：全局相似/互补，局部相似/互补。

5. 分析结果的展示

边缘概率分布（矩阵中的行或列）之间的关系可视化。

条件概率分布（矩阵块）的关系的可视化。

可以采用关系图，聚类，各种投影等方法。

6. 细节数据的验证

对原始数据的表格展示，分析师对案例的验证与评估。

下周工作：

毕业论文与专利

试图从企业数据中发现一些说服力较强的案例